HMMA 307 : Advanced Linear Modeling

Chapter 3 : ANOVA

COIFFIER OPHELIE GAIZI IBRAHIM LEFORT TANGUY

https://github.com/opheliecoiffier/CM_Anova

Université de Montpellier

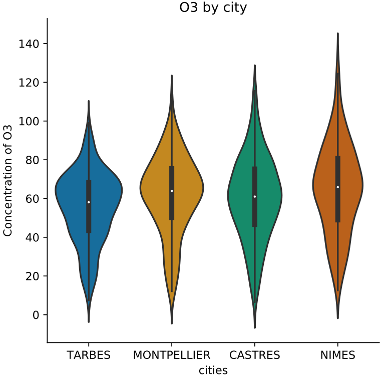
Statistical model for the ANOVA

ANOVA with the constraint $\sum \alpha_i^* = 0$

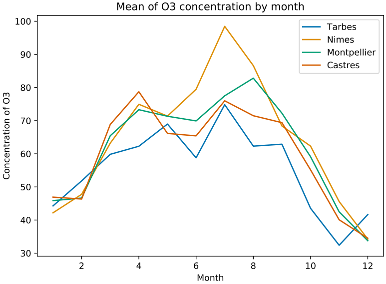
ANOVA with the constraint $\sum_{i=1}^{I} n_i \alpha_i = 0$

Non parametric alternative: permutation test

Comparison of the pollution between four cities



(a) Violin plot to compare the concentration of ozone between four cities in Occitanie.



(b) Mean of O3 by month for four cities.

Statistical model

Model equation

$$y_{ij} = \mu_i^* + \varepsilon_{ij}$$

- $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is the noise and $\operatorname{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \delta_{ii'} \delta_{jj'}$
- y_{ij} is the j^{th} measurement for that modality
- $\bar{y}_n$ is the average of y *i.e.*,

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}; i \in [\![1, I]\!].$$

Results from ANOVA and normality hypothesis

```
poll = ols('valeur_originale ~ C(nom_com)',data=df).fit()
sm.stats.anova_lm(poll, typ=2)
_, (__, ___, r) = sp.stats.probplot(poll.resid, fit=True)
```

Table: Results from the ANOVA on the O_3 concentration by cities.

	sum_sq	df	PR(>F)
C(nom_com)	16471.58	3	$3.86e^{-08}$

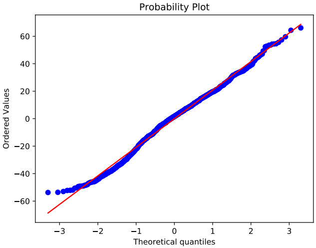


Figure: Check residues normality assumption

global/specific effect

We sometimes write : $\mu_i^* = \mu^* + \alpha_i^*$ to show the global mean effect and the specific effect of each feature.

Rem: With estimators $\hat{\mu}$ and $\hat{\alpha}_i$ for μ^* and α_i^* (for all $i = 1, \dots I$):

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

and

$$(\hat{\mu}_1, \dots, \hat{\mu}_I) \in \underset{(\mu_1, \dots, \mu_I) \in \mathbb{R}^I}{\arg\min} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

global/specific effect

We sometimes write : $\mu_i^* = \mu^* + \alpha_i^*$ to show the global mean effect and the specific effect of each feature.

Rem: With estimators $\hat{\mu}$ and $\hat{\alpha}_i$ for μ^* and α_i^* (for all $i = 1, \dots I$):

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

and

$$(\hat{\mu}_1, \dots, \hat{\mu}_I) \in \underset{(\mu_1,\dots,\mu_I)\in\mathbb{R}^I}{\arg\min} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

Thanks separability for $f(x_1, \dots, x_I) = \sum_i g_i(x_i)$

$$\min_{(x_1,\dots,x_I)} f(x_1, \dots, x_I) \iff \min_{x_i} g_i(x_i), \ i = 1, \dots, I$$

leading to

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i,:}.$$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then:

$$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.*, $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then:

$$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.*, $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$
- **associated estimator** :

$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_1}(y_{ij} - \mu - \alpha_i)^2$$

ANOVA : case of a modeling with : $\sum \alpha_i^* = 0$

Notice that if we change $\mu^* \longrightarrow \mu^* + \delta$ and $\alpha_i^* \longrightarrow \alpha_i^* - \delta$ then:

$$\mu_i^* = (\mu^* + \delta) + (\alpha_i^* - \delta)$$

- **hypothesis** : $\sum_{i=1}^{I} \alpha_i^* = 0$ *i.e.,* $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$
- **associated estimator** :
$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_1}(y_{ij} - \mu - \alpha_i)^2$$
- **Lagrangian** :
$$\mathcal{L}(\mu, \alpha, \lambda) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \mu - \alpha_i)^2 + \lambda \sum_{i=1}^{I} \alpha_i$$

Resolution of the optimization system

$$\nabla \mathcal{L}(\hat{\mu}, \hat{\alpha}, \hat{\lambda}) = 0$$

$$\begin{cases} \sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \hat{\mu}} = 0 \\ \frac{\partial \mathcal{L}}{\partial \hat{\alpha}_{i_0}} = 0,\ \forall i_0 \end{cases} \iff \begin{cases} \sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ n\hat{\mu} + \sum_{i=1}^{I} n_i \hat{\alpha}_i - n\bar{y}_n = 0 \\ n_{i_0}\hat{\mu} + n_{i_0}\hat{\alpha_{i_0}} = n_{i_0}\bar{y}_{i_0,:} - \hat{\lambda},\ \forall i_0 \end{cases}$$

$$\iff \begin{cases} \displaystyle\sum_{i=1}^{I} \hat{\alpha}_i = 0 \\ \hat{\mu} + \frac{1}{n} \sum_{i=1}^{I} n_i \hat{\alpha}_i = \bar{y}_n \\ n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + \hat{\lambda} = 0,\ \forall i_0 \end{cases}$$

Resolution of the optimization system

We have : $\sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + I\hat{\lambda} = 0$, so for $i_0 = 1, \cdots, I$,
so we get

Resolution of the optimization system

We have : $\sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:}) + I\hat{\lambda} = 0$, so for $i_0 = 1, \cdots, I$, so we get

$$\begin{aligned} & \sum_{i_0=1}^{I} n_{i_0}(\hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0}) + I\hat{\lambda} = 0 \\ \Longleftrightarrow \quad & n\hat{\mu} + \sum_{i_0=1}^{I} n_{i_0}\hat{\alpha}_{i_0} - \sum_{i_0=1}^{I} n_{i_0}\bar{y}_{i_0,:} + I\hat{\lambda} = 0 \\ \Longleftrightarrow \quad & n\hat{\mu} + \sum_{i_0=1}^{I} n_{i_0}\hat{\alpha}_{i_0} - n\bar{y}_n + I\hat{\lambda} = 0 \\ \Longleftrightarrow \quad & I\hat{\lambda} = 0 \Leftrightarrow \hat{\lambda} = 0 \end{aligned}$$

Results

- $\hat{\alpha}_{i_0} + \hat{\mu} = \bar{y}_{i_0,:}$
- $\hat{\mu} = \frac{1}{I} \sum_{i_0=1}^{I} \bar{y}_{i_0,:}$

Meaning that

$$\hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \frac{1}{I} \sum_{i_0=1}^{I} \bar{y}_{i_0,:}.$$

<u>Rem</u>:

- $\hat{\mu} \neq \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_n$
- It might be different if there are i, i' such that: $n_i \neq n_{i'}$

The weighted sum of the individual effects is zero

- **hypothesis** :

$$\sum_{i=1}^{I} n_i \alpha_i = 0$$

- **associated estimator** :

$$\underset{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I}{\arg\min} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

- **Lagrangian** :

$$\mathcal{L}(\mu, \alpha, \lambda) = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2 + \lambda \sum_{i=1}^{I} n_i \alpha_i$$

Resolution of the optimization system

$$\nabla \mathcal{L}(\hat{\mu}, \hat{\alpha}, \hat{\lambda}) = 0$$

$$\begin{cases} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \mu} = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_{i_0}} = 0 \ \forall i_0 \end{cases} \iff \begin{cases} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ n\hat{\mu} + \sum_{i=1}^{I} n_i \hat{\alpha}_i - n\bar{y}_n = 0 \\ \hat{\mu} + \hat{\alpha}_{i_0} - \bar{y}_{i_0,:} + \hat{\lambda} = 0, \forall i_0 \end{cases}$$

$$\iff \begin{cases} \sum_{i=1}^{I} n_i \hat{\alpha}_i = 0 \\ \hat{\mu} = \bar{y}_n \\ \hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \hat{\lambda} - \bar{y}_n, \forall i_0 \end{cases}$$

Results

- We multiply the third line of the equation by n_{i_0} then we add them up for i_0 in 1 to I. We finally obtain $\hat{\lambda} = 0$,
- $\hat{\mu} = \bar{y}_n$

Meaning that:

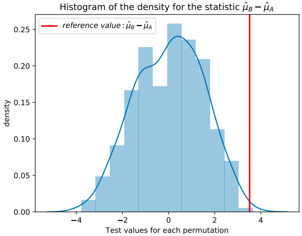
$$\hat{\alpha}_{i_0} = \bar{y}_{i_0,:} - \bar{y}_n.$$

Rem: The next case to study will be: $\alpha_{i_0} = 0$

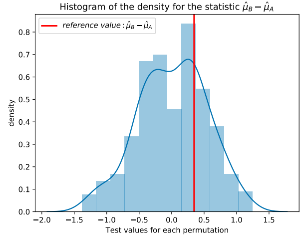
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,



$\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

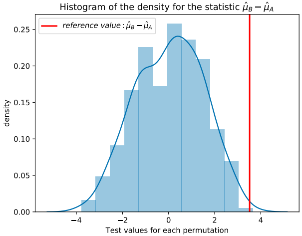


$\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

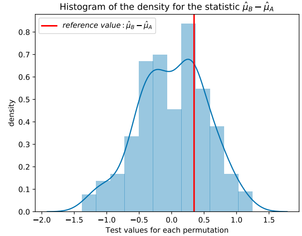
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- H_0: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),



$\mu_A^* = 3$, $\mu_B^* = 7$, we reject the equality.

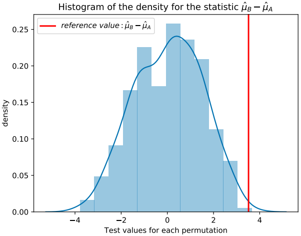


$\mu_A^* = 2$, $\mu_B^* = 2.5$, we don't reject the equality.

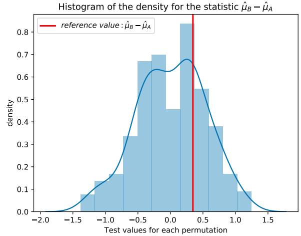
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- H_0: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,



$\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

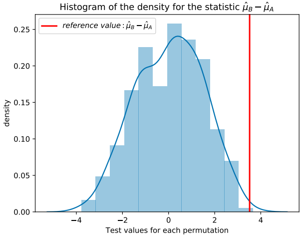


$\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

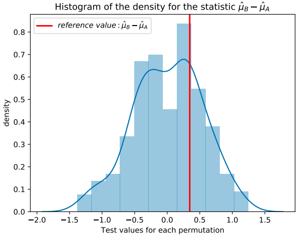
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- H_0: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,



$\mu_A^* = 3,\ \mu_B^* = 7$, we reject the equality.

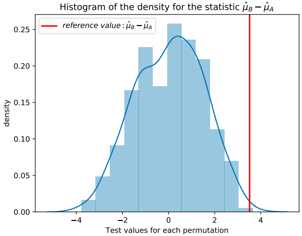


$\mu_A^* = 2,\ \mu_B^* = 2.5$, we don't reject the equality.

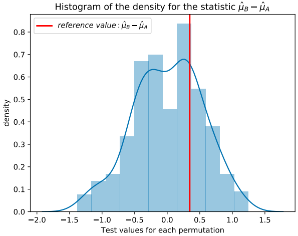
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- H_0: $\mu_A^* \geq \mu_B^*$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,
- shuffle the groups and recalculate the test statistic J times,



$\mu_A^* = 3$, $\mu_B^* = 7$, we reject the equality.

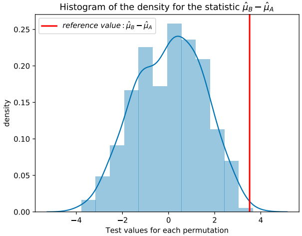


$\mu_A^* = 2$, $\mu_B^* = 2.5$, we don't reject the equality.

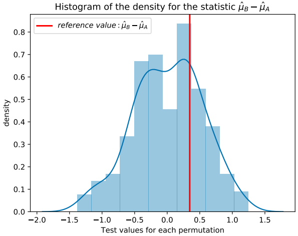
Permutation test: medical scenario

Protocol (Monte-Carlo):

- 2 groups: A the control and B the test, we test the effect of the treatment,
- H_0: $\mu^*_A \geq \mu^*_B$ (Test if the treatment is better),
- Assign values for the effect of the treatment,
- Get the reference statistic: $\hat{\mu}_B - \hat{\mu}_A$,
- shuffle the groups and recalculate the test statistic J times,
- p-value is the number of statistics over the reference divided by J.



$\mu^*_A = 3,\ \mu^*_B = 7$, we reject the equality.



$\mu^*_A = 2,\ \mu^*_B = 2.5$, we don't reject the equality.

Case: $\alpha_{i_0} = 0$

Our 3 hypotheses:

- $\sum_{i=1}^{I} \alpha_u = 0$
- $\sum_{i=1}^{I} \alpha_i x_i = 0$
- $\alpha_{i_0} = 0$

Case: $\alpha_{i_0} = 0$

Our 3 hypotheses:

- $\sum_{i=1}^{I} \alpha_u = 0$
- $\sum_{i=1}^{I} \alpha_i x_i = 0$
- $\alpha_{i_0} = 0$

Associated estimator:

$$\min_{(\mu,\alpha)\in\mathbb{R}\times\mathbb{R}^I} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{n} (\mu + \alpha_i - y_{i,j})^2$$

$$\mathcal{L}(\mu, \alpha, \lambda) = \sum_{i=1}^{I} \sum_{j=1}^{n} (\mu + \alpha_i - y_{i,j})^2 + \lambda \alpha_{i_0}$$

Case: $\alpha_{i_0} = 0$

- $i \neq i_0 : \frac{\partial \mathcal{L}}{\partial \alpha_i} = \sum_{i=1}^{n_i} [\hat{\mu} + \hat{\alpha}_i - y_{i,j}] = 0 \quad (*)$
- $i = i_0 : \frac{\partial \mathcal{L}}{\partial \alpha_{i_0}} = \sum_{i=1}^{n_i} [\hat{\mu} + \hat{\alpha}_i - y_{i,j}] + \hat{\lambda} = 0 \quad (**)$
- $\hat{\mu} = y_{i_0,j} - \hat{\lambda}$

Case: $\alpha_{i_0} = 0$

$$\sum_{i \neq i_0} (*) + (**) = \sum_{i \neq i_0} \sum_{j=1}^{n_{i_0}} \hat{\mu} + \sum_{j=1}^{n_{i_0}} \hat{\mu} + \sum_{i \neq i_0} \hat{\alpha}_i + n_{i_0} \hat{\alpha}_{i_0} - \sum_{i \neq i_0} \sum_j y_{i,j} \tag{1}$$

$$- \sum_{j=1}^{n_{i_0}} y_{i,j} \tag{2}$$

$$= \sum_i \sum_j \hat{\mu} + \sum_i n_i \hat{\alpha}_i - \sum_i \sum_j y_{i,j} + \hat{\lambda} \tag{3}$$

$$= 0 \tag{4}$$

Case: $\alpha_{i_0} = 0$

$$\sum_{i \neq i_0} n_i \hat{\mu} + \sum_{i \neq i_0} n_i \hat{\alpha}_i - \sum_{i \neq i_0} \sum_{j} y_{i,j} = 0$$

With the previous equation:

$$n_{i_0} \hat{\mu} + n_{i_0} \hat{\alpha}_{i_0} - \sum_{j=1}^{n_{i_0}} y_{i,j} + \hat{\lambda} = 0$$

$$\implies \hat{\mu} + \hat{\alpha}_] i_0 - \bar{y}_{i_0} + \frac{\hat{\lambda}}{n_{i_0}}$$

$$\implies \hat{\mu} = \bar{y}_{i,:} - \frac{\hat{\lambda}}{n_{i_0}}$$

Case: $\alpha_{i_0} = 0$

$$n_i(\bar{y}_{i_0,:} - \frac{\hat{\lambda}}{n_{i_0}}) + n_i\hat{\alpha}_i - n_i\bar{y}_{i,:} = 0$$

$$\implies \hat{\alpha}_i = \frac{\hat{\lambda}}{n_{i_0}} - y_{i_0,:} \bar{+} y\bar{_{i,:}}$$

We admit that $\hat{\lambda} = 0$

$$\implies \begin{cases} \hat{\alpha}_i = \bar{y}_{i,:} - \bar{y}_{i_0,:} \\ \hat{\alpha}_{i_0} = 0 \\ \hat{\mu} = \bar{y}_{i_0,:} \\ \frac{\partial \mathcal{L}}{\partial \hat{\alpha}_{i_0}} = 0 \ \forall i_0 \end{cases}$$

Variance estimator

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=i}^{n_i}(\bar{y}_{i,:} -_{i,j})^2$$

- $n - I$: Correction so that $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$
- $y_{i,j} = \mu^* + \varepsilon_{i,j}$
- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$

Variance estimator

Notice :

$X = [\mathbb{1}_{C_1}, \dots, \mathbb{1}_{C_I}] \in \mathbb{R}^{n \times I}$:

$$\frac{1}{n - rg(X)} \left\| y - X\hat{\beta}^{LS} \right\|^2 \text{ unbiased estimator of } \sigma^2$$

$\sum_{i=1}^{I} \mathbb{1}_{C_i} = \mathbb{1}_n \; rg(\tilde{X}) = I, \tilde{X} = [\mathbb{1}_n, \mathbb{1}_{C_1}, \dots, \mathbb{1}_{C_I}]$

where the C_i are the indexes of observations of the i^{th} category

Test: "are the effect all the same?"

The null hypothesis: H_0

$$H_0 : \mu_1^* = \mu_2^* = \cdots = \mu_I^*$$

- $F_{obs} = \frac{\frac{1}{I-1}\sum_{i=1}^{I}(\bar{y}_{i,:} - \bar{y}_n)^2}{\hat{\sigma}^2}$ with: $F_{obs} \sim \tilde{F}_{n-I}^{I-1}$
- We reject the test: $F_{obs} > F_{n-I}^{I-1}(1-\alpha)$ (if we want to test α)

Bibliography

- Salmon, Joseph. *Modèle linéaire avancé : Anova*. 2019. URL: http://josephsalmon.eu/enseignement/Montpellier/HMMA307/Anova.pdf.
- Wilber, Jared. *Monte-Carlo method (permutation test)*. 2019. URL: https://www.jwilber.me/permutationtest/.